Characterization of Salinity Impact on Synthetic Floc Strength Via Nonlinear Component Analysis

Hang Yin[1], Patrick Carriere[1], Huey Lawson[1], Habib Mohamadian[1], Zhengmao Ye[1]

[1]College of Science and Engineering, Southern University, Baton Rouge, LA 70813, USA
{hang_yin, patrick_carriere, huey_lawson, habib_mohamadian, zhengmao_ye}@subr.edu

Abstract. Many complex mechanisms are inherently engaged in flocculation processes with nonlinear nature. The strength of synthetic flocculates or natural flocculates may be relevant to numerous factors as well. It will be expensive and virtually impossible to determine an exact influential list among various factors via trial and error experiments exclusively. The objective is to develop an analytical scheme for decision making about the relevant influential list at least cost. Multivariate statistical methods are actually capable of differentiating dominating factors. There is no existing research outcome being documented about applications of either principal component analysis (PCA) or nonlinear component analysis (NCA) to the whole area of flocculation and coagulation research, essentially optimization has been never achieved indeed. Compared with PCA, NCA is more versatile to solve large dimensional nonlinear multi-variate problems with a potential to reach infinite dimensionality. NCA is thus proposed in a preliminary study to figure out feasibility of challenging research to extract dominating factors associated with the mechanical behavior of flocs. Without convincing evidence so far on specific utmost factor in the floc strength studies, the scale of adjustable salinity has been intentionally chosen as the first principal component to interpret variations observed in the simulation results, together with interconnections to other major principal components. Based on the pioneering methodology proposed, some interesting results are well obtained and documented. At the same time, there is no technical difficulty unquestionably to extend the proposed NCA approach to multivariable and high-dimensional nonlinear cases.

Keywords: Flocculation; Flocculate Strength; Salinity; Principal Component Analysis (PCA); Nonlinear Component Analysis (NCA)

1 Introduction

Flocculation and coagulation are two major processes for solid and liquid separation. The floc arises from the flocculation process when fine particulates are gradually accumulated to clump together. There are a vast variety of factors affecting floc strength, including salinity, pH scale (acidic, alkaline, neutral), particles (concentration, size and turbidity), mineral precipitates (e.g. aluminium salt, iron salt, aluminium or iron hydroxide), biological matters (e.g., bacteria, fungi, virus), organic substances

(e.g. humus, oils), geometries (e.g. surface porosity, roughness, microstructure, nanostructure), physical properties (e.g. temperature, pressure, humidity), hydro-dynamics (e.g. velocity variability, flowrate), and so on. Dilute suspensions could also give rise to colloidal dispersions and interactions with remarkably complex nature. In fact it leads to a complicated nonlinear multivariate analysis problem [1-2]. Among these factors, there is no doubt that the salinity effect is one of commanding factors being involved. Its impact on the synthetic floc strength will be analyzed as a typical case study.

In order to experimentally test the strength of flocs and mineral, some state-of-the-art techniques have been adopted to observe the deformation mechanism under various load levels even at nanoscale. For instance, the effect of loading on the nanoscale deformation modes has been investigated under repeated nanoindentation loading on muscovite with a sharp indenter tip, so as to analyze the deformation mechanisms at nanoscale on a basis of hardness and elastic modulus normal to the basal plane. The testing curves show nonlinear characteristics upon loading and unloading such as the closed hysteresis loops. The transition from the high Young modulus to low bulk modulus occurs due to 3D confinement surrounding an indenter tip in the plastic shakedown process [3-4].

To avoid or to minimize laborious, expensive, risky and time-consuming processes in the relevant civil engineering studies, and to be away from instrument-specific, field-specific and technology-specific conditions, the multivariate model could serve as a promising solution. It ranges from the classical multivariate regression model to popular principal component analysis (PCA), as well as to more powerful nonlinear component analysis (NCA) and independent component analysis (ICA). Even though no application of PCA on mechanical behaviors of the floc strength has appeared in literatures, various cases of successful research have been conducted across related fields. For instance, failures occur frequently at wastewater treatment sites, which cause terrible environmental implications. The reliable and versatile PCA helps to group soils independently of classifications to distinguish the appropriate soils for sustainable long-term effluent irrigation and to locate influential parameters for actual characterization [5]. PCA is also proposed for sensor fault detection to monitor the structure health such as cracks on the underground structure. The useful results could be extended to micro-crack detection of the concrete [6]. The decision-making of drought quantification depends on various statistical aspects. The multivariate PCA technique has been applied to hydrological drought monitoring. A typical Streamflow Drought Index (SCI) has been expanded to multivariate index at multiple time scales. In general the first principal component can be used to interpret majority of regional variations [7]. Reliable data acquisition of soil permeability in fact is vital for soil-water research. An accurate prediction model will be beneficial to identify those key soil parameters without high cost and length time needed. From PCA analysis on over 90 samples examined with 16 parameters at 37 sites, five variables are determined to be of strong correlation with soil permeability in a preliminary study [8]. PCA can be also applied to characterization of biomedical samples. In Raman spectroscopic study, PCA has been carried out to differentiate diverse tissue samples using scatter plots together with artificial intelligence techniques [9].

PCA always provides convincing outcomes for linear feature extraction. Limitation of PCA however lies in data analysis of nonlinear high dimensional spaces. Especially for nonlinear problems in high dimensional or even infinite dimensional cases, PCA itself could be sometimes vulnerable. As an alternative, with the involvement of a nonlinear kernel, kernel PCA is capable of extracting nonlinear features. NCA has the priori advantage with the possibility to extract more principal components than linear PCA. NCA classification also works pretty well with a relatively limited number of principal components compared with feature space dimensionality. To each the same classification performance, much fewer nonlinear principal components are needed than those in the linear case [10]. Another ICA approach is also powerful for nonlinear cases. It covers the steps of centering, whitening and independent optimization. ICA has the merit of minimizing statistical dependence of all components. It has been applied to spatial object recognition problems in remote sensing areas [11].

2 Instrumentation and Data Normalization

All natural floc samples are directly collected from the Atchafalaya Bay in Gulf of Mexico. The natural floc and synthetic floc samples are both tested at Lab using a UTM (Universal Testing Machine) together with a patent licensed compression cell. The dynamic nature of cohesive sediments has a strong impact on the mechanical properties of clay flocs in aqueous environment. Testing samples have been prepared with aqueous clay suspensions and instant ocean salt solution for floc generation. The applied suspension has a constant clay concentration of 0.4 g/L. The flocculation of suspended clay particles occurs in the stirring bath of the Particle Size Analyzer (PSA). The clay to clay collision continues upon stirring. Three targeting salinities for synthetic flocs are 2, 10 and 30 PSU respectively, while the natural floc acts as the sample reference instead. For each clay flocs group, over a dozen individual flocs are examined to provide a sufficient sample population for further statistical analysis. A sketch of instrumentation and experimental setup is shown in Fig. 1.

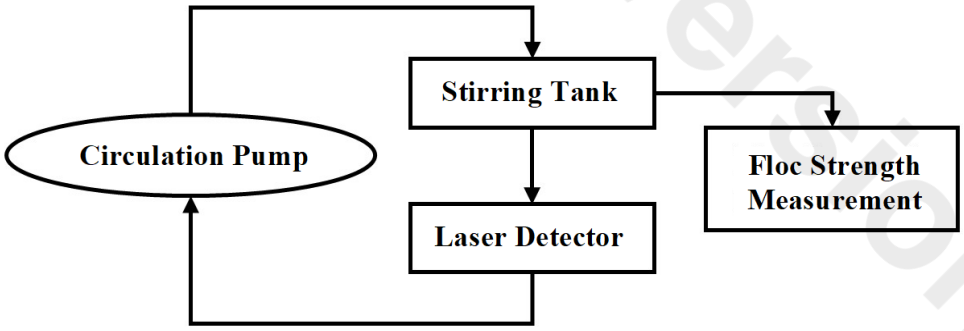

Fig. 1. Sketch of Flocculation Generation Systems

In Fig. 1, the stirring tank is used for sample preparation. Synthetic floc samples are prepared inside by mixing the Na-mont suspension with the guar gum in water with different salinities. Using a Laser detector, when formed flocs are traveling along the circulation pipes, the particle size is measured and plotted until floc size distribution is determined to be stable. The circulation pump forces water inside to continuously flow in order to avoid sedimentation. Then selected individual floc samples will be collected which are transported through a pipette to a universal testing machine for floc strength measurement. For each test, the compressive load and displacement are both recorded for analysis subsequently to determine breakage strength. Rather than the costly nano-scale contact mechanics based approach being developed, promising nonlinear component analysis has been employed for the first time to analyze the floc compression curves in this study. A typical testing result for each clay floc group is plotted in Fig. 2, representing a sufficient set of sample population. To conduct the further statistical analysis, data normalization is necessary.

$$x_i = \frac{z_i - \bar{z}}{s} \tag{1}$$

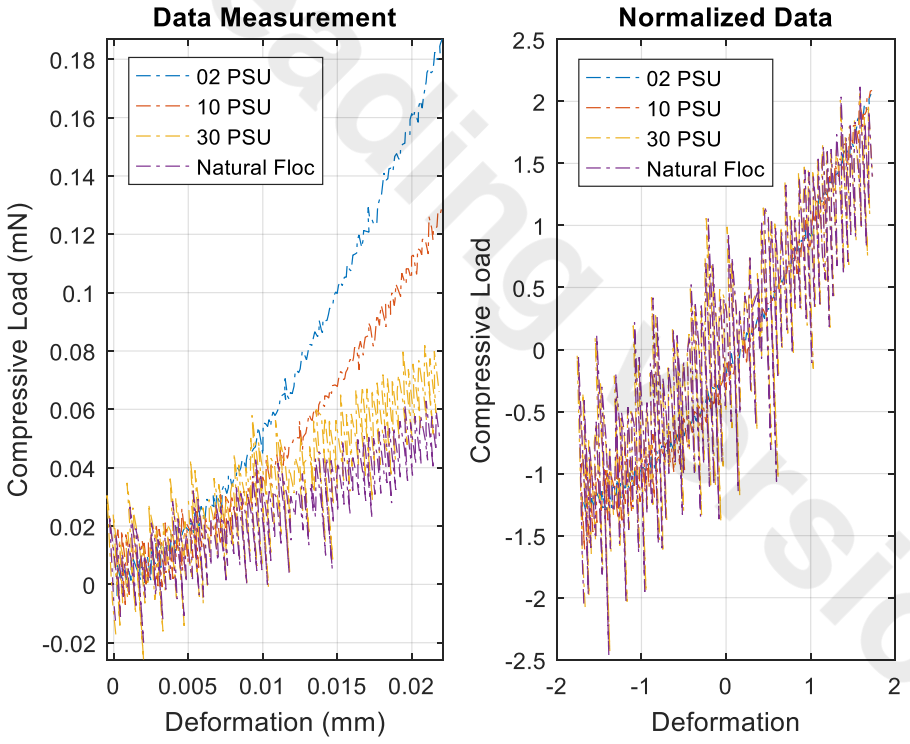

Fig. 2. Test Data and Data Normalization

The normalized data are computed as a ratio of difference between the i^{th} observation data and the sample mean to the sample standard deviation based on all conducted

measurements, as shown in (1) where z_i refers to the i^{th} set of observation data and x_i refers to the i^{th} set of normalized data; z_bar refers to the sample mean and S refers to the sample standard deviation. The typical normalized data are also shown in Fig. 2. From Figs. 1-2, among four sets of data, the compressive load-deformation curve of the natural floc (roughly 36 PSU) collected from the Gulf of Mexico is much less deterministic than those synthetic flocs of 2, 10 and 30 PSU. The randomness reflects that highly nonlinear mechanical behaviors occur in the floc strength study.

3 Nonlinear Component Analysis

PCA is unsupervised numerical dimensionality reduction technique to examine the underlying variability of multi-dimensional data. PCA generates the orthogonal transformation of the coordinate system for visualization of experimental data in terms of sample arrays. Some handful principal components are sufficient to describe complex data structure in new axes. Data rotation is implemented to maximize the variance. High dimensional data are projected into a low dimensional subspace. The underlying variables of experimental data could be extracted by solving eigenvalue problems [9]. For a set of N centered observations x_i (i=1, 2, …, N) in the feature space, PCA in fact implements diagonalization for the covariance matrix C_{PCA} in (2).

$$C_{PCA} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} x_i x_i^T \tag{2}$$

It is conducted by solving the eigenvalue equation (3):

$$\lambda v = C_{PCA} v = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} (x_i \cdot v) x_i \tag{3}$$

where λ is an eigenvalue of C_{PCA} and v is the correspondent eigenvector. x_i is the centered data and C_{PCA} is a positive definite matrix in general. The solutions lie in a span of $x_1, x_2, \ldots, x_N,$, which is formulated in (4).

$$\lambda(x_i * v) = (x_i * C_{PCA} v) \text{ for } i=1, 2, \ldots, N \tag{4}$$

For high-dimensional feature extraction of the nonlinear behaviors (e.g. mechanical properties) with respect to the actual flocculation processes, principal components of various features are in fact nonlinearly related to the complex mechanical behavior itself. The nonlinear kernel approach is proposed to deal with this tough engineering issue for the first time. The nonlinear kernel PCA could extract the more substantial features than the linear PCA for both classification and prediction purposes, in typical nonlinear feature space of high dimensionality and even up to infinite dimensionality. With the centered data ($\sum_{i=1}^{N} \Phi(x_i) = 0$), i=1, 2, …, N, NCA is able to implement diagonalization on the covariance matrix C_{NCA} in (5).

$$C_{NCA} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \Phi(x_i) \Phi(x_i)^T \tag{5}$$

The eigenvalue equation (6) is now to be solved.

$$\lambda V = C_{NCA}V = \frac{1}{N}\sum_{i=1}^{N}(\Phi(x_i)\cdot V)\Phi(x_i) \tag{6}$$

where λ is an eigenvalue of C_{NCA} and V is the correspondent eigenvector. $\Phi(x_i)$ is the centered data and C_{NCA} is a positive definite matrix in general. Solutions lie in a span $\Phi(x_i)$, $\Phi(x_2)$, …, $\Phi(x_N)$ where Φ defines a nonlinear function. The nonlinear kernel function is defined as an inner product in the feature space being denoted as (7).

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \tag{7}$$

NCA has been applied to project data in a higher dimension space to lower dimension space without necessity of explicit mapping. It can even reach infinite-dimensional cases. Without loss of generality, three typical nonlinear kernel functions are selected: Gaussian kernel, Laplace kernel and Cauchy kernel, respectively. The three kernel functions are formulated as (8-10).

$$k_G(x, y) = e^{-\|x - y\|^2 / 2\sigma^2} \tag{8}$$

The Gaussian kernel is an exponential kernel in which sigma acts as an adjustable parameter to be set. It depends on the tradeoff between nonlinearity and sensitivity.

$$k_L(x, y) = e^{-\|x - y\| / \sigma} \tag{9}$$

The Laplace kernel is another exponential kernel in fact. It can suppress sensitivity against the sigma parameter variations.

$$k_C(x, y) = \frac{1}{1 + \|x - y\|^2 / \sigma^2} \tag{10}$$

The Cauchy kernel instead follows the Cauchy distribution. It provides the long-range influence and sensitivity over the high dimensional space. The focus of the article is however to demonstrate feasibility of new NCA engineering practices on the synthetic floc strength study. Details on comparing merits and drawbacks of three typical non-linear kernels will be discussed in another subsequent article.

4 Numerical Simulations

Numerical simulation results based on the NCA approach are depicted in Fig. 2 to Fig. 7, respectively. The 1st nonlinear principal component (PC1) always represents a coordinate direction with the greatest variation, which is of the maximal variance. The 2nd nonlinear principal component (PC2) represents a direction with the maximal variation still remained in the data which is orthogonal to PC1. The PC3 is orthogonal to both PC1 and PC2, the PC4 is orthogonal to PC1, PC2 and PC3, and so on.

In Fig. 3, the scores (eigenvalues) corresponding to the first 10 nonlinear principal components via kernel NCA have been shown, representing the 10 latent nonlinear factors with strongest correlation to mechanical properties, such as the salinity, pH scale, organic substance, biological matter, mineral precipitate, geometry, and so on. For the natural floc curve tested, the variance of PC1 to PC10 drops gradually but no PC is exceptionally high indicating that the strength of the natural floc depends on a

number of factors instead of individual one. On the other hand, up to 6 PCs (PC1 to PC6) could contribute to over 60% of the total data variation. Thus the emphasis should be put on six prevailing factors for the floc strength analysis. It is however unconvincing to claim the superiority list of these factors without the scientific proof. The proposed approach is to adjust the scale of salinity in the synthetic flocs (2 PSU, 10 PSU, 30 PSU) while the natural floc (36 PSU) serves as the reference. PC1 always contributes the most to the variation. Apparently additional variations on PC1 of synthetic flocs (2 PSU, 10 PSU, 30 PSU) come from major changes on salinity, as no other factors differ from those of the natural floc (36 PSU). The largest difference on the scale of salinity also gives rise to the highest mismatch on PC1 between synthetic floc (2 PSU) and natural floc (36 PSU). The PC curves of two synthetic flocs (2 PSU and 10 PSU) are similar while the curve of the synthetic floc (30 PSU) resembles that of the natural floc (36 PSU). Based on data of PC1 to PC5, the curves become sharper when the scale of salinity moves further away from the reference level of the natural floc due to the superior role of salinity on the mechanical behavior. Influential factors might also be intrinsically coupled together, variations of some other PCs could occur accordingly. It depicts evidently that NCA could be successfully applied to the nonlinear floc strength analysis. When changes on some other influential factors are also involved besides the salinity, an identical NCA methodology could be applied while several PCs will be accessed simultaneously to determine the exact priority list among all the corresponding influential factors.

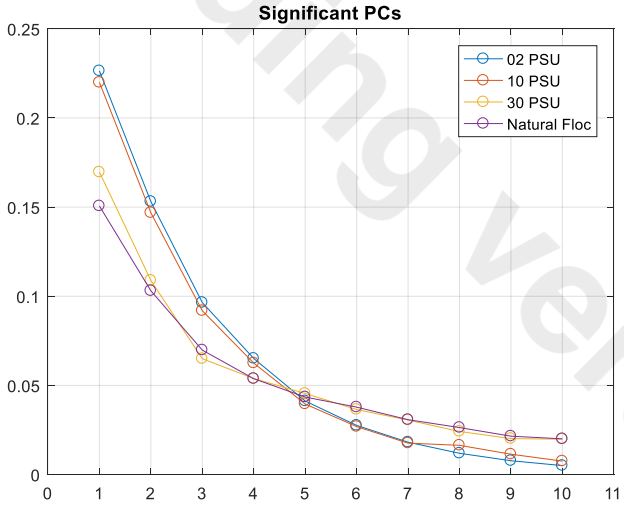

Fig. 3. First 10 Principal Components Via Nonlinear Kernel PCA

The NCA study also reveals internal correlations among multiple PCs. A simple way is to examine 3 dominating PCs in the 3D plots. Despite of the fact that the nonlinear kernel type selection is out of scope of this work, which acts as another part of future research, the Gaussian kernel, Laplace kernel and Cauchy kernel have all been used

so that some differences could be visually observed. In Fig. 4, 3D results (PC1 vs PC2 vs PC3) based on Gaussian kernel and Laplace kernel are quite similar but the 3D result based on Cauchy kernel exhibits a remarkable difference. Higher deviation on scale of salinity being artificial tuned (2 PSU) from that of the natural floc leads to more deterministic curve. The smaller mismatch (30 PSU) leads to more stochastic pattern similar to that of the natural floc. From PC1 to PC3, the variation level will decrease step by step.

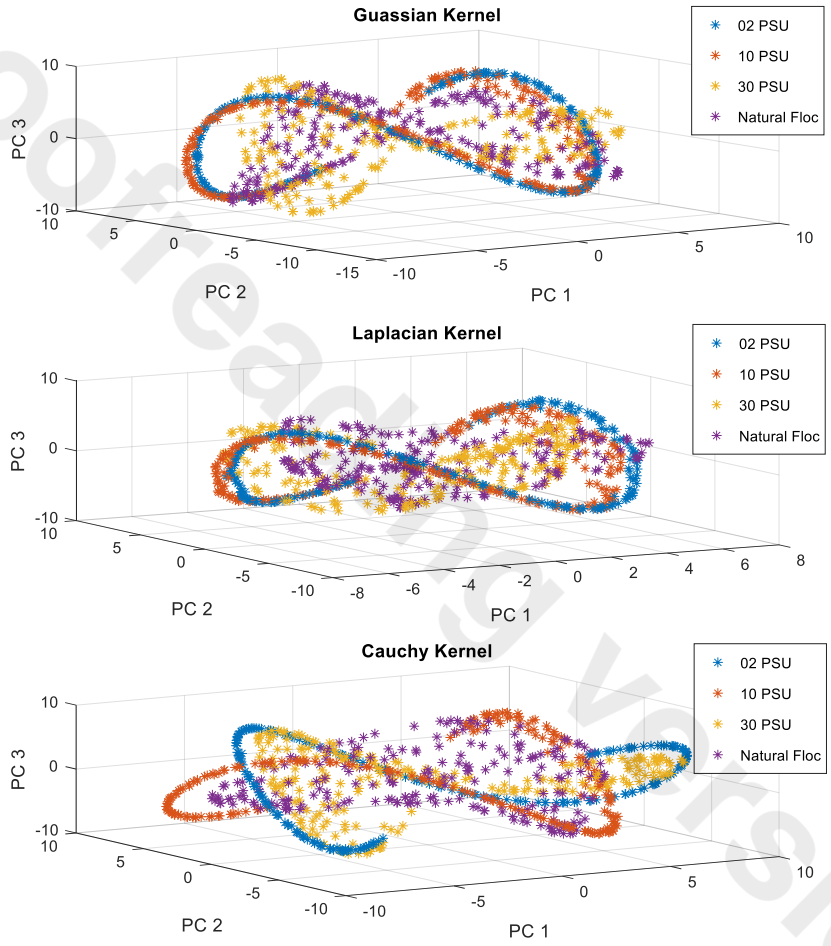

Fig. 4. Linearized 3 Dominant Principal Components

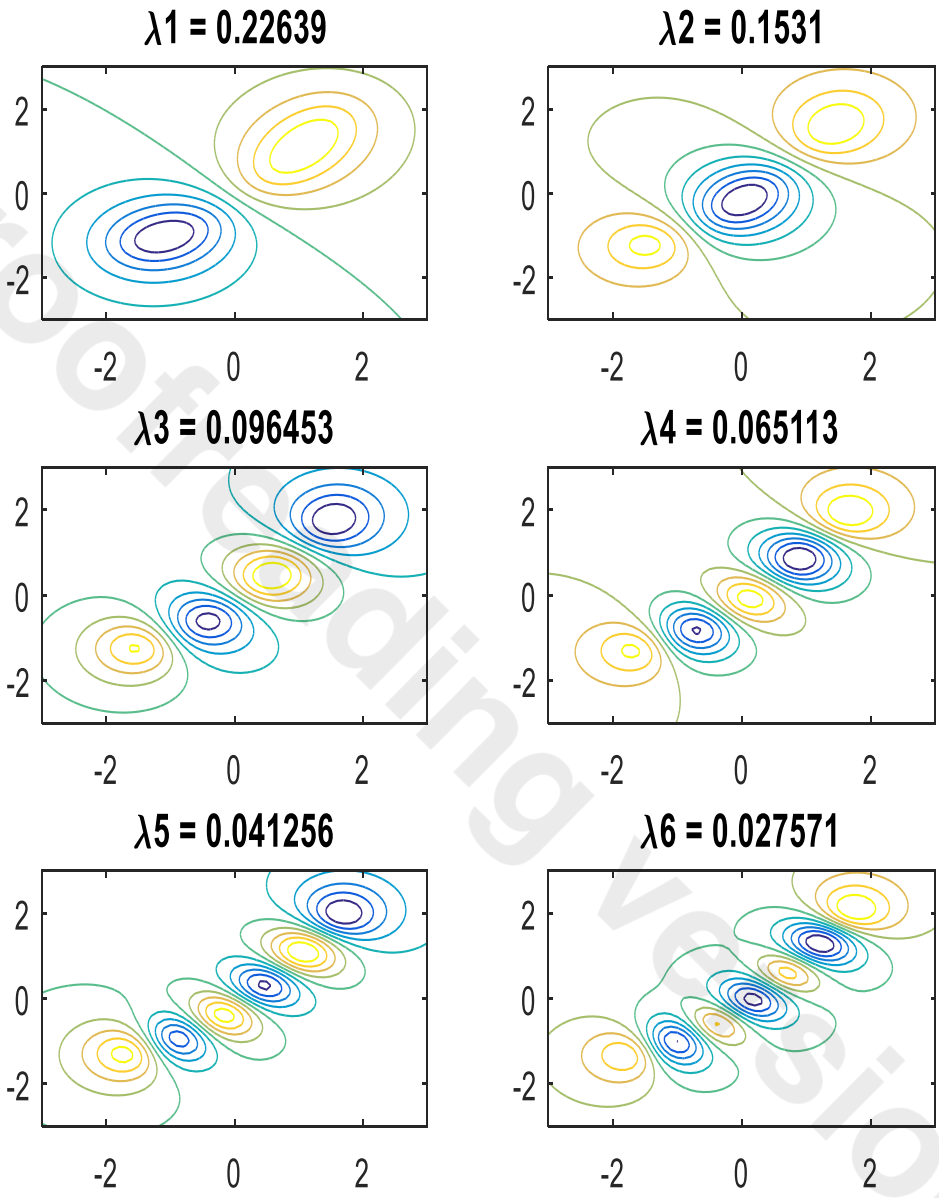

Fig. 5. Projection 2D Contours of Eigenvectors (PCs) – 2 PSU

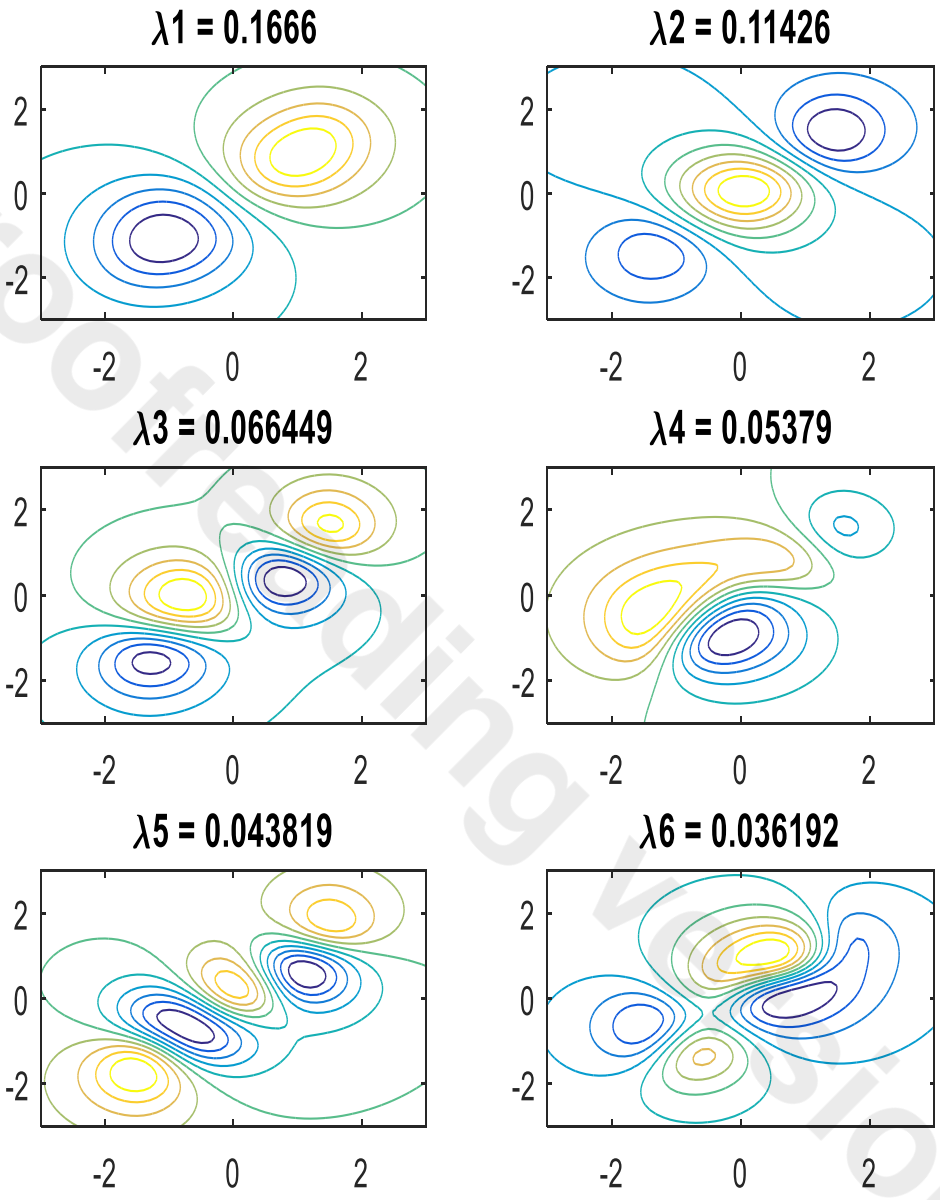

Fig. 6. Projection 2D Contours of Eigenvectors – Natural Floc

For PC1 to PC6 instead, corresponding to individual constant significant eigenvalues, contour curves can be plotted which represent projection of the principal component vectors onto the 2D plane in the linear space. These contour curves are orthogonal to the corresponding principal component (eigenvector) in fact. For example, contour curves for the synthetic floc (2 PSU) are shown in Fig. 5 and contour curves for the natural floc (36 PSU) are shown in Fig. 6. The unique variations of patterns provide another evidence of the salinity impact on the floc strength.

5 Scatter Plot Analysis

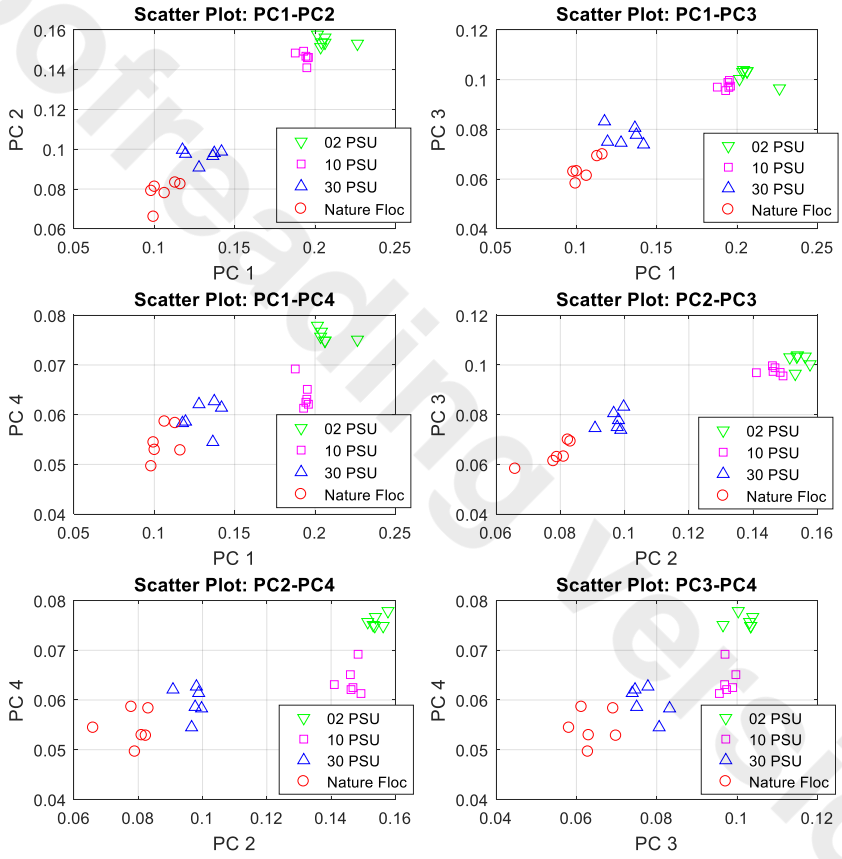

Fig. 7. Nonlinear Component Analysis - Scatter Plots

Classification and prediction are two potential and challenging engineering practices for all the research, especially when at least 5 or 6 influential factors are all taken into account. Some interesting results have been obtained by comparing PC1 to PC4 in the

2D scatter plots. The simulation data arise from diverse sets of experimental data at several scales of salinity. The scatter plot is shown in Fig. 7. The first 4 leading PC scores (eigenvalues) are plotted in pair, such as PC1 vs PC2, PC1 vs PC3, and PC3 vs PC4. It could be extended to more complex cases of arbitrary number of factors with no extra technical effort at all. From Fig. 7, four sets of data of synthetic flocs (2 PSU, 10 PSU, 30 PSU) and natural floc (36 PSU) can be distinguished clearly in each of 6 plots. Each set of data could be grouped as a single cluster. The distance between any 2 clusters also reflect the mismatch on the scale of salinity. With large amount of data, classification can be easily made using NCA and scatter plots immediately. Meanwhile without the priori hypotheses needed, some future on-site sample data can be analyzed directly and quickly via the proposed NCA approach, possibly in real time, so that the range of influential factors (e.g. salinity) will be determined for accurate model prediction instantly.

6 Conclusions

A reliable nonlinear multivariate approach has been presented to characterize the deformation mechanism of flocculation in this preliminary study, where the reachable scale of salinity has been deliberately used as the first principal component on the synthetic floc strength research. Comparisons on mechanical properties between the natural floc strength and synthetic floc strength at different salinity scales have been made. Among a plenty of variables that actually affect the mechanical property of flocs, a number of controlling factors with strong correlation should be captured for classification and redundancy removal. Experimental methods are obviously lengthy, burdensome and costly. Nonlinear component analysis instead has the capability of significant dimensionality reduction for nonlinear problems in the high dimensional space. Thus some powerful parameters in the low dimensional linear space are used to manifest the physical phenomena occurred in the high dimensional nonlinear space. Impact of salinity scale variations on the floc strength has been explored which has verified the feasibility of the NCA methodology. Interesting results from scatter plots are helpful for both classification and prediction purposes potentially. The proposed approach can be extended to the highly nonlinear cases at the large dimensional space in a straightforward way.

References

1. Shainberg, I. and Levy, G., "Flocculation and Dispersion", Encyclopedia of Soils in the Environment, pp. 27-34, 2005, Elsevier Ltd
2. Theng, B. K. G., “Formation and Properties of Clay-Polymer Complexes”, Volume 4, 2nd Edition, pp. 511, 2012, Elsevier, Amsterdam
3. Yin, H., and Zhang, G. (2011), “Nanoindentation Behavior of Muscovite Subjected to Repeated Loading”, ASME Journal of Nanomechanics and Micromechanics, 1(2), 72–83
4. Yin, H., and Zhang, G. (2011), “Cyclic Nanoindentation Shakedown of Muscovite and Its Elastic Modulus Measurement”, Proceedings of the Society for Experimental Mechanics

Series, Volume 4, 2011, MEMS and Nanotechnology, Proceedings of Annual Conference on Experimental and Applied Mechanics, 2011, Springer, New York, 83–92

5. Dawes, L., and Goonetilleke, A., (2006), “Using Multivariate Analysis to Predict the Behaviour of Soils Under Effluent Irrigation”, Water, Air and Soil Pollution 172 (1-4): pp. 109-127. Springer
6. Stoffels, N., Sircoulomb, V., Hermand, G., Hoblos, G., "Principal Component Analysis for Fault Detection and Structure Health Monitoring", pp. 1751-1758, 7th European Workshop on Structural Health Monitoring, July 8-11, 2014, Nantes, France
7. Arabzadeh, R., Kholoosi, M., Bazrafshan, J., "Regional Hydrological Drought Monitoring Using Principal Components Analysis", ASCE Journal of Irrigation and Drainage Engineering, v.142, n.1, 20 pp., JAN 2016
8. Yulianti, M., Sudriani, Y. and Rustini, H., "Preliminary Study of Soil Permeability Properties Using Principal Component Analysis", IOP Conference Series: Earth and Environmental Science, pp. 1-5, Volume 118, 2018
9. Ye, Z., “Artificial Intelligence Approach for Biomedical Sample Characterization Using Raman Spectroscopy”, IEEE Transactions on Automation Science and Engineering, Vol. 2, No. 1, pp. 67-73, January 2005
10. Schölkopf, B., Smola, A. and Müller, K., "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", pp.1299-1319, Neural Computation, Volume 10, Issue 5, July 1, 1998, MIT Press
11. Ye Z., Mohamadian H. and Ye Y., "Independent Component Analysis for Spatial Object Recognition with Applications of Information Theory", Proceedings of IEEE World Congress on Computational Intelligence, pp. 3640-3645, Hong Kong, June 1-6, 2008